Review

Low-cost evaluation techniques for information retrieval systems: A review

Shiva Imani Moghadi, Sri Devi Ravana, Sudharshan N. Raman

A Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia
b Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Malaysia

Abstract

For a system-based information retrieval evaluation, test collection model still remains as a costly task. Producing relevance judgments is an expensive, time consuming task which has to be performed by human assessors. It is not viable to assess the relevancy of every single document in a corpus against each topic for a large collection. In an experimental-based environment, partial judgment on the basis of a pooling method is created to substitute a complete assessment of documents for relevancy. Due to the increasing number of documents, topics, and retrieval systems, the need to perform low-cost evaluations while obtaining reliable results is essential. Researchers are seeking techniques to reduce the costs of experimental IR evaluation process by the means of reducing the number of relevance judgments to be performed or even eliminating them while still obtaining reliable results. In this paper, various state-of-the-art approaches in performing low-cost retrieval evaluation are discussed under each of the following categories; selecting the best sets of documents to be judged; calculating evaluation measures, both, robust to incomplete judgments; statistical inference of evaluation metrics; inference of judgments on relevance, query selection; techniques to test the reliability of the evaluation and reusability of the constructed collections; and other alternative methods to pooling. This paper is intended to link the reader to the corpus of ‘must read’ papers in the area of low-cost evaluation of IR systems.

Contents

1. Introduction .................................................. 302
2. High cost relevance judgments .................................. 303
   2.1. Relevance judgment and pooling method .................. 303
   2.2. Pooling effects ........................................ 305
3. Performing low-cost evaluations ................................ 306
   3.1. Evaluation measures robust to incomplete judgment .......... 306
   3.2. Selecting best set of documents for judgment .......... 307
   3.3. Statistical inference of evaluation metrics .................. 308
   3.4. Inference of relevance judgment .......................... 308
   3.5. Topic selection ....................................... 308
   3.6. Techniques to construct reliable and reusable test collections .... 309
   3.7. Alternative methods to pooling ........................ 310
4. Conclusion .................................................. 311
Acknowledgment ............................................. 311
References .................................................. 311
1. Introduction

Performance characteristics of retrieval systems deals with the accuracy of produced results which are about how effective an information retrieval (IR) system is in retrieving relevant documents. In order to evaluate the effectiveness of IR systems, two different approaches that may complement each other can be adopted. These are: user-based and system-based methods. The user-based approach concentrates on observing the user’s interactions with the system to quantify their satisfaction levels (Fidel, 1993). This method deals with obtaining and analyzing the user’s feedbacks on the retrieval performance, user interface and other aspects of the system. The user-based method requires lots of human participation, and indeed will be costly and time consuming.

On the other hand, the system-based retrieval evaluation focuses on experiments that are aimed to evaluate the performance of the retrieval algorithm and considers users as an abstraction (Mandl, 2008). Such evaluation usually utilizes a test collection (Sanderson, 2010). The test collection consists of a document corpus, queries and set of judgments on relevance, which will be elaborated in Section 2.1 (Baeza-Yates & Ribeiro-Neto, 1999; Mandl, 2008; Melucci & Baeza-Yates, 2011). On the other hand, system scores are basically computed based on a chosen evaluation metrics. An evaluation metric quantifies the similarity between the set of documents retrieved by the systems (also known as runs) and a set of relevant documents (qrels) to see how good a retrieval system is. Due to repeatability, reusability and scalability characteristics of this system-based experiment, it provides a proper environment for evaluation and performing experimental experiences that makes this approach to be important (Baeza-Yates & Ribeiro-Neto, 1999).

Performing IR evaluation through the test collections is costly since part of this method relies on human effort. Assessing the relevancy of documents to a topic is a time consuming, and expensive task that has to be performed by human assessors who are usually specialists in one or more areas of knowledge. Due to the huge number of documents in the document corpus (to simulate the real world search engines), having a complete relevance judgment in a collection such as TREC is not viable. TREC is an initiative by the National Institute of Standards and Technology (NIST) and U.S. Department of Defence which provides the necessary common platform for research within the IR community for large-scale evaluation of retrieval methods. Table 1 shows the number and size of documents in the TREC document corpus, and number of topics in various TREC experiments from 1998 to 2010. For example, for TREC-2010 Web, to obtain a complete judgment set, a total of 1 billion documents should be assessed by the experts. Judging 1 billion documents with limited number of experts is nearly impossible as it will incur high cost in terms of judgment effort (hiring of suitable assessors to represent real users) and time-consumed during the judgment process. For example, by assuming that two documents could be successfully judged by an assessor within a minute (see Section 6.3, Sanderson, 2010), it will take about 347,000 days (or 950 years) to judge 1 billion documents. In order to judge more documents within a limited time, more assessors need to be hired which would also incur cost. This example shows that while the cost for relevance assessment could be easily quantified, there are other issues that need to be considered as well to ensure consistency in judgments by assessors and generate reliable evaluation results.

The pooling method was proposed by Spárk Jones and van Rijsbergen (1975), and adopted by subsequent IR evaluation initiatives in order to decrease the number of judgments that need to be created. In this method, a set of top d ranked documents returned by participating systems (in the TREC experiment) are selected to create the pool of documents that need to be judged. Then, all the duplicate documents are eliminated from the pool and followed by a judgment for relevancy by the assessors. The judgment set generated is called the partial relevance judgments because not all documents from the corpus were used for the assessment. All the documents outside the pool were considered as non-relevant. TREC was the first initiative that used partial relevance judgment based on a pooling method as a substitution for complete judgment (Spárk Jones & van Rijsbergen, 1975). Some other initiatives for experimental based evaluation such as Cross Language Evaluation Forum (CLEF), Forum for Information Retrieval Evaluation (FIRE), NACIS Test Collection for Information Retrieval (NCTIR), Chinese Web Information Retrieval Forum (CWIRF), and the IR Initiative for Evaluation of XML retrieval (INEX) are also using similar methods to generate relevance judgment sets.

Using this partial relevance judgment set, participating systems can be fairly evaluated provided that the systems have contributed to the pool. Problem arises when new systems or improved systems need to be evaluated using this same relevance set. In the experimental based evaluation, reusing of the relevance judgments set is practiced because generating a new set of relevance judgments will incur additional time and effort. This practice of reusing judgments is difficult and
develops a possibility of the judgments to be biased against the new or improved system because the documents retrieved by these systems may not have been pooled for judgment by assessors. However, it is common to have both types of systems (systems contributing to the pool and systems that do not contribute to the pool) to be evaluated using the same set of relevance judgments in TREC like initiatives. Hence, it is important to be able to obtain evaluation results that are highly accurate and reliable.

As shown in Table 1, in order to mimic the Web search which involves millions of documents, the size of the corpus of documents is increased in retrieval experiments. In turn, there is a need to perform low-cost evaluations while still obtaining reliable results. Since IR researchers are seeking for solutions to reduce the high cost of judgments on relevance, various low-cost evaluation methods have been proposed and examined.

This paper discusses on the issues associated with high cost relevance judgments in retrieval experimentation and presents a state-of-the-art review on alternative techniques available to perform low-cost systems evaluation. Here, evaluation cost can be defined as expenses (in terms of hiring expert judges and time spent) in assessing the judgments for relevance. Hence, the difference between a low-cost and high-cost system evaluation can be represented as the number of hours or days spent by expert judges in assessing judgments (in generating a reliable judgment set) besides the salary paid for their effort. The coverage of this paper is distinct from existing review papers on IR evaluation as it presents the compilation of various low-cost evaluation techniques that can be undertaken in an experimental based IR system evaluation. To the best knowledge of the authors, at least to date, a review paper such as this is not available although the techniques pointed out in this review have been discussed in various other papers individually and not as a compilation. It is intended to link researchers to the corpus of ‘must read’ papers in the area of low-cost evaluation of IR systems.

This introductory section establishes the reason why low-cost IR evaluation methodologies are essential. The subsequent section highlights the processes involved in creating relevance judgments and undertaking the pooling method. This is then followed by issues with the judgment process that makes it a costly task. In the subsequent section, recently proposed methods on performing low-cost IR evaluation will be reviewed in terms of its advantages and how the presented techniques can save experimental cost. The conclusion section presents a future of IR evaluation and some suggestions of research areas that can be explored further.

2. High cost relevance judgments

As an overview of the evaluation process and how it relates to incurring high costs in retrieval experiments, issues on generating relevance judgments based on a pooling method are discussed in this section.

2.1. Relevance judgment and pooling method

Using a test collection model is a common system-centered approach adopted for IR evaluation. This approach is also referred to as the Cranfield method based on the Cranfield tests carried out between 1957 and 1966 which is the origins of today’s laboratory retrieval evaluation experiments (Cleverdon, 1967). As mentioned earlier, a test collection framework consists of: (i) a corpus of documents which is a collection of a great number of documents; (ii) topics that are usually a set of pre-defined queries formed in a standard format and will be used by a retrieval system as search query (queries will be referred to as topics in the rest of this paper); and (iii) relevance judgment set which is made by human assessors commonly generated based on a pooling method of the retrieved documents. The pooling method has been adopted in IR evaluation experiments since it produces sufficient number of judgments for achieving reliable results (Voorhees, 1998). The relevance assessment scale may be in a binary or graded format. A binary judgment means a retrieved document is either relevant or not relevant. Graded judgments are such as, highly relevant, relevant and non-relevant documents. These judgments will be tagged with the value of 2, 1 and 0 respectively in the judgment set. The TREC-9 Web track and TREC-2005 Enterprise track are some examples of a three-point relevance scale (Hawking, 2000). However, Sormunen (2002) argued that the distinction between grades in the three-point relevance scale (used in the TREC-9 Web track) is not clearly defined. A four-point relevance scale for IR evaluation experiments was proposed where relevance criteria were defined so that a distinction is made between relevant and highly relevant documents and marginally relevant documents (Sormunen, 2002). Note that the need for graded relevance was realized due to the nature of some IR systems, where it is important to retrieve highly relevant documents and not just relevant documents. Whether an experiment should use a binary, three-point relevance scale, four-point relevance scale or multi-point relevance scale is all dependent on the purpose and need of the IR experiments.

Fig. 1 shows the processes involved in a typical IR evaluation experiment based on the test collection model. The test collection (represented by the three elements within the dotted boundary) consists of document corpus, topics and relevance judgments. As shown in the figure, systems seeking to be evaluated (System A to System F) examine their retrieval algorithms on a test collection and return a set of documents (called runs) as the result of their search. For instance in TREC experiments, each system submits a set of at least one and no more than 1000 documents (in most cases) in decreasing relevance to the topic as its run.

On the other hand, systems that are participating in the pool construction are considered as contributing systems (System A to System D) and others as non-contributing (System E and System F). As mentioned previously, partial judgment is performed by selecting a set of top d ranked documents retrieved for each topic from the runs (by contributing systems),
and pool of documents is created for assessment. For instance, a subset of \( d = 50 \) to \( d = 100 \) of top ranked documents for each topic, retrieved by System A, B, C and D will be selected for the pool. All the documents in the pool are then judged by human assessors to recognize their relevancy to that specific topic. Thus relevance judgment set or qrels are created based on assessors’ decisions. All the unjudged documents outside the pool are considered as non-relevant.

When the judgment set is ready (refer to the tick and cross symbols within the dotted boundary), the whole set of runs that are returned by retrieval systems, contributing to the pool or not, will be compared against it. The retrieval effectiveness of systems are measured based on the number, fraction and relevance of the returned results to the topic. System performance is quantified according to these factors using a chosen evaluation metrics such as Average Precision or Normalized Discounted Cumulative Gain (NDCG). Per-system per-topic scores are calculated for systems and scores are aggregated to obtain a single overall performance score of a system on a set of topics. Common score aggregation techniques to combine per-system per-topic scores is the arithmetic mean. Other aggregation techniques such as the geometric mean, harmonic mean and median were also explored (Ravana & Moffat, 2009). According to the calculated scores, systems will be ranked in decreasing order of the quality of their performance. On the whole, the systems ranking is generated for all the systems as a result of each IR evaluation experiment.

Several evaluation metrics can be used to measure system performance. An evaluation metric can be defined as a function that takes an ordered vector of relevance values, and returns a single numeric score, that summarizes those values (Webber, 2010). Recall and precision are two fundamental measures used for the retrieval evaluation and many other evaluation metrics are generated based on this concept. Metrics such as Precision-at-ten (P@10), Average Precision (AP), Mean Average Precision (MAP), R-precision and NDCG (Järvelin & Kekäläinen, 2002) have been commonly used to compute per-system per-topic scores. Other metrics such as standardized average precision (SP) (Webber, Moffat, & Zobel, 2008), binary preference
(bpref) (Buckley & Voorhees, 2004), ranked biased precision (RBP) (Moffat & Zobel, 2008) and many others were proposed in order to overcome the disadvantages in the commonly used metrics.

2.2. Pooling effects

As described earlier, relevance judgment causes a weak point in the system based evaluation because this approach is still dependant on human efforts. Accordingly, the characteristics of performing relevance assessments makes it an expensive and time consuming task, since human assessors have to spend several hours to read each document and define its relevancy to a specific topic. On the other hand, because of the judgment's specifications, assessors may become bored and less precise, which may affect the accuracy of the judgment (Melucci & Baeza-Yates, 2011). Moreover, assessors may disagree in declaring a document as relevant or not. Many issues can influence human judgment and may make it to be dependent on the assessor's personality (Harter, 1996). Previous experiments have shown that sometimes the same assessor did make errors in identifying the relevancy of similar documents (Bernstein & Zobel, 2005). Even a previously relevant document may be considered as non-relevant after a period of time because the assessment task is human dependant (Sanderson, Scholer, & Turpin, 2010). In addition to the effect of time passing between judging two relevant documents on the assessor’s choice, recent investigations have shown consistency in judging relevant documents will influence the decision about the relevancy of documents that follow (Scholer, Turpin, & Sanderson, 2011). On the whole, assessors’ faults in judging may have great impact on the system rankings (Carterette & Soboroff, 2010).

Within the IR research community, researchers are not only looking for alternatives in terms of reducing the experimental cost, but are continuously working on identifying alternative methods that would eliminate or neutralize human subjectivity. However, this paper mainly focuses on methods or techniques that contribute to low-cost IR evaluation experiments in terms of higher accuracy and reliability of evaluation results without additional cost for preparing judgments. Note that, some of the discussed techniques may have also been proposed to address the issue of human subjectivity as well as to the relevance judgment process.

Even though the pooling technique is adopted by evaluation initiatives to decrease the number of judgments needed to be performed, it has some drawbacks. The disadvantages of the pooling method according to Moffat, Webber, and Zobel (2007) are that all the systems contributing to the pool would not be equally accurate and the existence of defective systems may affect the pooling method. So, a pool of judgments may suffer from containing irrelevant documents that causes the assessors to evaluate thousands of irrelevant items. Another issue is the insensitivity of pooling. This means that all systems, contributing to the pool or otherwise, would be assumed as equal and scored with the same level of reliability. The third issue is that pooling does not consider the hardness (difficulties) of topics since some topics may require more number of judged documents. For example, TREC-2009 Web track consisted of 50 topics from 450 to 500 (Clarke, Craswell, & Soboroff, 2009). As shown in Fig. 2, some topics have more relevant documents in the test collection in comparison with other assessed documents. This indicates that the retrieval systems may find relevant documents on some topics much easier than others.

Although using the pooling method decreased the number of judgments, there are still large numbers of documents to be judged (Zobel, 1998). As an example, Table 2 shows the number of documents and judgments produced for TREC-2004 Robust
track and TREC-2009 Web track (Clarke et al., 2009; Voorhees, 2004). This indicates that even when documents are pooled, for TREC-2004 Robust track, assessors will still need to spend several days to judge more than 300,000 of documents (which is about 58.9% of the total number of documents in the corpus). Besides, the fast growth of the number of documents and topics due to high speed Web developments and accessibilities make the relevance assessment process even a more complex and costly task. Rapid changes in science and technology increase the necessity of comparing different IR systems. Growth in number of documents, topics, retrieval systems may all result in more judgments or in other words additional costs to the evaluation process. On the whole, all these issues encourage IR researchers to seek methods to perform evaluation with fewer judgments and at the same time keep the experimental results reliable. In addition to the stated issues, the relevance judgment based on a pooling method may cause a biased evaluation results for new systems that did not contribute to the pool. This means that although the pooling method have been commonly used, some systems may be unfairly ranked since only a fraction of documents will be judged in this method (Buckley, Dimmick, Soboroff, & Voorhees, 2007). In contrary, efforts can focus on building reliable test collections with more number of topics while performing fewer judgments rather than assessing large amounts of documents (Sanderson & Zobel, 2005).

### 3. Performing low-cost evaluations

Since the number of documents in test collections are growing fast to simulate the Web, researchers are looking for approaches that could facilitate in keeping up with this speed (Mandl, 2008). As a result, IR evaluation methods for performing low-cost evaluations in addition to achieving reliable results have been designed. Different approaches in performing low-cost retrieval evaluations can be categorized under these titles as discussed by Carterette, Kanoulas, and Yilmaz (2010):

- i. calculating evaluation measures robust enough to cater for incomplete judgments
- ii. selecting the best sets of documents to be judged
- iii. statistical inference of evaluation metrics
- iv. inference of relevance judgments
- v. query selection
- vi. techniques to test the reliability of the evaluation and reusability of the constructed collections, and
- vii. alternative methods to pooling.

#### 3.1. Evaluation measures robust to incomplete judgment

This section discusses evaluation measures that are effective in the presence of a limited number of judgments due to pooling. Based on the experiments that will be discussed below, some metrics can be used to calculate more accurate and reliable results while there is less effort for making relevance judgments. Buckley and Voorhees (2004) expressed the need for a proper evaluation measure for large collections and introduced a measure, $b_{pref}$, to calculate system scores. Because of the rapid growth in the number of documents especially in dynamic environments such as Web, relevance judgment sets may be unsatisfactory in addition to be incomplete. Buckley and Voorhees (2004) examined these effects on IR evaluation experiments by computing system rankings for some TREC collections repeatedly while keeping the same number of judgments and reducing the number of documents. The results demonstrated that some measures were not consistent to massively incomplete judgments and also was not a good choice for calculating system scores. This was the reason they proposed $b_{pref}$ as a preference based measure which calculates systems score based only on judged documents. The $b_{pref}$ measure considers top non-relevant documents and tries to solve the problem of partial judgment by neglecting the search results for which relevance data is not available. If we consider a topic with $R$ judged relevant documents, $N$ judged non-relevant documents, $r$ as a relevant document retrieved and $n$ number of judged non-relevant documents retrieved by the system, Eq. (1) calculates $b_{pref}$ as (Buckley & Voorhees, 2004):

$$
\text{b}_{\text{pref}} = \frac{1}{R} \sum_{r} \left( 1 - \frac{n \text{ ranked } \text{ higher than } r}{\min(R, N)} \right)
$$

In contrast to most measures, the $b_{pref}$ measure is related to the proportion of non-relevant documents that are retrieved before the relevant ones. Comparison between system rankings produced using $b_{pref}$, MAP, $P_{10}$ or $R$-Precision showed that the ranking based on $b_{pref}$ were more consistent when using 100% of qrels. While reducing the size of qrels, $b_{pref}$ still remained reliable until the size was cut down to half.
On the other hand, Sakai (2007) proposed applying some metrics such as Q-measures, NDCG and AP to a condensed lists, gained from removing all unjudged documents from the initial ranked lists, in order to obtain more accurate results. Since bpref was proposed as an appropriate measure in the case of having a large collection with incomplete and imperfect judgments, Sakai (2007) compared it with this proposed method. The experiment revealed that Q-measures, NDCG or AP can produce even more accurate results than bpref when these are applied to a condensed list. In order to examine the accuracy of system rankings, Sakai (2007) investigated his proposed method in four different test collections from NTCIR initiatives. Here, Q-measures and NDCG were suggested as better alternatives to measure the performance of systems in the presence of incomplete judgments. In terms of low-cost IR evaluations, the higher ranking accuracy of these evaluation metrics would not require more relevance judgments and eventually the effort for assessing more judgments could be avoided (Sakai, 2007).

Ahlgren and Gronqvist (2008) conducted an experiment focusing on the measures dependent on the number of relevant documents using data from TREC collections. The aim was to examine reliability and stability of system rankings under different levels of relevance of data incompleteness. Comparing bpref and rankEff with MAP, with regard to the number of relevant documents, was to identify their behavior in situations where there are not many relevant documents in the pool. Based on the findings, when the amount of relevance judgments is low, the metric rankEff is a more stable measure in comparison with the other two metrics. The metric RankEff as mentioned in Ahlgren and Gronqvist (2006, 2008) was inspired by bpref-10. If I is the set of all known non-relevant documents for topic t, r is the number of relevant documents and n the total number of relevance judged documents, RankEff for the IR method M and topic t is

$$\text{RankEff}(M, t) = \frac{\sum_{d \in I} l(d)}{r(n-r)}$$

where l(d) is the number of non-relevant documents d such that d \in I. Since, this metric has the ability to generate evaluations measures to rank systems relatively in the same order under different levels judgment incompleteness, it is a reliable metrics to be applied in a low-cost IR evaluation experiment especially when the judgments are less.

### 3.2. Selecting best set of documents for judgment

Selecting the best set of documents to be judged is an important issue in generating reliable and accurate IR evaluation because it will result in producing a reliable judgment set besides eliminating additional costs. This section discusses different approaches in selecting documents for creating relevance judgments. Some approaches focus on increasing the reliability of experimental results in creating judgment sets, while others focus on reducing the number of judgments that need to be made to decrease the experimental costs.

Zobel (1998) introduced a method to judge a shallow pool of documents. Subsequently, based on the evaluation results and the shallow pool, the judgment pool can be extended using more documents. In this technique, relevant documents are recognized faster than the traditional pooling method. In the same year Cormack, Palmer, and Clarke (1998) proposed the Move-To-Front (MTF) pooling method. This method assumes that higher ranked documents and also documents retrieved by systems that have been recently recognized as a relevant document are more likely to be relevant. Based on these assumptions, a priority order will be set in the pool. They showed that by using this approach, a judgment set can be constructed faster. Even though the number of relevant documents in this approach was approximately the same as the available judgment set, it took less time and human effort to be created.

Carterette and Allan (2005) proposed a method to create test collections and relevance judgments in an incremental process. The algorithm would select documents for judgment from the corpus of documents intelligently and stops judging when it considers that no more judgments are needed (Carterette & Allan, 2005). It selects documents that seem to contribute more information about the differences in calculating MAP, and ends the process when there is a significant difference (statistically). This incremental method requires less number of judgments to select documents to be judged compared to the classic pooling method. This method can be used by researchers to create an in-house test collection which allows them to perform IR evaluation on small sets of systems. A weight is given to each document using Eq. (3), where \( r_i(d) \) is the rank of document \( d \) in ranked list \( i \).

$$w_d = \left| \frac{1}{r_1(d)} - \frac{1}{r_2(d)} \right|$$

Apart from experiments that focused on forming test collections, many researchers have also concentrated on determining the order or the number of documents to be assessed. Moffat et al. (2007) proposed a method to adjust the order that documents are added to a queue for judgment. In their approach, static and dynamic approaches were defined to select documents that can produce a more accurate judgment set. In the static approach, documents for judgment were chosen based on their value in the scoring rule. In other words, if a document was retrieved by most of the participating systems, it would be considered as a higher rank for judgment. On the other hand, in the dynamic approach, the results of relevance judgments that has been made previously, has had impacts on selecting the following documents for judgment.

In another experiment, Carterette, Allan, and Sitaraman (2006) proposed an algorithm based on AP measures to select documents for judgment. This method could select documents to be judged in a minimum time and produce reliable system rankings. The confidence in AP results can be estimated since it is dispersed over relevance judgment sets. This can be used
as a feasible state to stop the judgment for test collections. Whilst this approach selects documents to be judged in a minimal time, it can be appropriate when little relevance judgments are available in the evaluation environment.

### 3.3. Statistical inference of evaluation metrics

As a solution to the problem of large scaled retrieval evaluation, Aslam, Pavlu, and Yilmaz (2006) proposed a statistical method to estimate standard evaluation measures such as AP and R-precision based on random sampling. They discussed that the existing methods may generate biased inferences of evaluation measures when there are few judgments available. In order to calculate more reliable scores while having limited number of judgments the researchers suggested a sampling method and tested it on TREC data collections. The findings indicated that if the number of sample increases there would be less variance in the score estimation. According to their findings, most judgments should be done close to top ranked documents which may have higher sampling probabilities (Aslam et al., 2006; Aslam & Yilmaz, 2007).

On the other hand, research in the field of performance evaluation usually focuses on measuring new or improved systems, meaning systems that did not contribute to the pool of judgments. Consequently, reusing of the system performance measures that were generated at the end of each evaluation experiment have not been considered (Ravana, 2011). Calculating smoothed evaluation measures was proposed by Ravana, Park, and Moffat (2009) to estimate more of the refined system scores without performing additional relevance judgments. To examine this method, TREC-2004 Robust track and its results have been used. The findings showed that blending partial previous information with new observations would result in calculating more stable system orderings from extended experiments. The experimentation showed that every-system evaluation can be carried out as a follow-on to initial small-scale experiments (Ravana et al., 2009). The evaluation metrics used were the AP, SP, P@10 and NDCG. According to this method, smoothed average precision for a topic set \( T \), a topic \( t \in T \), and system \( s \), is calculated as shown in Eq. (4). Smoothed average precision is computed using AP for system \( s \) and topic \( t \) (additional topic and it is not a subset of \( T \)) and MAP of the same system on topic set \( T \) (prior results). Here, \( \alpha \) is constant value that can vary between 0 and 1.

\[
\text{smoothed } \text{AP}_{s,t} = \alpha \text{AP}_{s,t} + (1 - \alpha) \text{MAP}_{s,T} \tag{4}
\]

### 3.4. Inference of relevance judgment

Identifying relevant documents amongst unjudged documents using existing judgment set, is another approach to decrease the evaluation’s costs. Aslam and Yilmaz (2007) proposed a method that would recognize the relevance of unjudged documents based on a small number of judged documents. In this way, pools of documents can be entirely judged with less human effort. The method was proposed to build judgment sets with real relevance assessments using smaller number of judgments than the traditional 100-depth pools. At first only some documents will be judged to attain the sampling estimates. Then an optimization procedure will be followed to obtain the inferred judgment set. Finally, only the documents that are marked as relevant in the inferred judgment set will be judged. The inferred relevance judgments were very similar to the actual available judgment sets and could reliably be used as the actual judgments.

On the other hand, Butcher, Clarke, and Yeung (2007) proposed methods to automatically make an unbiased judgment set from the primary, biased judgments achieved out of pooling. When using existing test collections to evaluate new systems, systems that did not contribute to the pool, relevance judgment set may be biased against them in comparison to contributing systems. They suggested that a more reliable result can be achieved if unbiased sets of judgments that predict the relevancy of unjudged documents are built. Whether an unjudged document is relevant or not was determined by using text classifiers (Butcher et al., 2007). A comparison between the results obtained from the proposed method and the produced results by bpref demonstrated that it may generate more accurate system scores especially if the original judgment is quite reliable.

### 3.5. Topic selection

How topics were developed for test collections and the numbers of topics used for IR evaluation experiment are issues that may have impacts on producing the judgment sets. Many researchers studied the impacts of the number of topics in a test collection when performing IR evaluation. The problem of high number of topics while having few judgments was addressed by Million Query Track at TREC which used 25 million documents and 10,000 topics. In this experiment, Carterette, Pavlu, Kanoulas, Aslam, and Allan (2008) discussed the issue that evaluation experiments with more topics and fewer number of judgments are more reliable, while assessors’ efforts can be reduced. In contrast, evaluation over fewer topics with more judgments may be less accurate. So it can be concluded that as there are more number of topics in the test collection, the possibility to produce more accurate system rankings is higher.

In studies conducted by Pal, Mitra, and Kamps (2010) to investigate the reliability of INEX test collections (another TREC like initiative but focusing on XML documents) by progressively reducing the pool size, they received similar results as Carterette et al. (2008). Their findings showed that to build consistent collections with fewer efforts, it is better to judge shallower pools for more topics rather than considering fewer number of topics. They concluded that to utilize available human resource efficiently, pool depth should be chosen according to the characteristics of each topic.
In a more recent experiment, Hosseini, Cox, Milic-Frayling, Vinay, and Sweeting (2011) proposed a method to develop the relevance judgment set for new systems. In view of the huge topics and document numbers in the creation of test collection, they suggested instead of spending the whole budget assessing all the topics in collection, just select a subset of topics and perform relevance judgment for those selected. They used a convex optimization strategy to prioritize queries.

3.6. Techniques to construct reliable and reusable test collections

While constructing test collections is an expensive and time consuming process, it is important to build collections that are reliable and also reusable for further experiments which will indeed reduce the cost and time of performing evaluations. In the case of reusability of test collections, some algorithms have been proposed and examined in order to create collections that are more cost effective. For instance, Carterette, Gabrilovich, Josifovski, and Metzler (2010) introduced the idea of estimating reusability of test collections while concentrating on measuring how useful existing judgments are for evaluating new systems. Two different methods were proposed and examined on the TREC collections for reusability measurement.

In the first technique, confidence intervals were generated for evaluation metrics on the basis of logistic regression. Using this method, existing relevance judgments are appropriate for evaluation of new systems if the interval is tight. However, when the interval is wide, more documents are required to be judged to evaluate the systems reliably. The other method changes standard evaluation metrics into reusable measures. Confidence intervals for effectiveness metrics such as MAP and P@k can be calculated when the relevance of unjudged documents is estimated. Based on the experiments on TREC data, the calculated confidence intervals were reliable and the measures were suitable for many interval tasks (Carterette, Gabrilovich, et al., 2010). In a different experiment, Carterette, Kanoulas, Pavlu, and Fang (2010) proposed an experimental design that gathers proof about whether a test collection is reusable or not, after relevance judgments were prepared. By growth in the size of the test collections, and, while producing relevance judgments were more costly, pooling may become less feasible. So, evaluation experiments can become cost effective while creating large test collections that are usable. In this approach, the method of selecting judgments were not important but they have tried to use more topics in addition to fewer numbers of judgments that can be an effective way of building test collections. This method provides a framework to recognize whether a test collection can be reused while the relevance judgment set is small (Carterette, Gabrilovich, et al., 2010).

Sanderson et al. (2010) investigated the behavior of human assessors in judging documents. Since decision making may differ with time, they examined the similarity of judgments that have been made in a sequence of time. Thus, they computed the distance between random pairs of relevant and non-relevant documents for some TREC test collections and tested if assessor’s view of relevance has changed across time. As the result of their investigation, they concluded if a series of similar documents were shown to assessors sequentially, their judgment would be more consistent. The consistency will reduce when the pairs are judged further apart. Consequently, to generate more reliable qrels, documents in the pool have to be clustered in proper order. This will assist assessors in making better decisions to produce consistent judgment sets that can be reused in future experiments.

Meanwhile, Webber and Park (2009) proposed a method to estimate the degree of bias against non-contributing documents and adjusted the system scores based on that. This method, based on a simple robust sampling was to correct bias affected systems. In this approach all the systems, contributing and non-contributing to the pool, would be evaluated against a common set of topics. Bias detected against non-contributing systems will be used to correct system scores on existing topics. This method is beneficial because of its robustness to biased judgments whereas it uses the least inferential assumptions to produce relevance judgments.

In a recent research, to improve the accuracy of evaluating contributing systems to the pool and certify that test collection is reliable for a new system, not contributing to the pool, Hosseini, Cox, Milic-Frayling, Sweeting, and Vinay (2011) proposed a two staged procedure. They first assign a fixed budget B1 to determine the relevance judgment set across all topics. Then by using information from the first step, acquired the prioritized topic-document pairs and assigning budget B2 to expand the relevance judgment set for non-contributing systems. In addition, Hosseini, Cox, and Milic-Frayling (2011) also proposed a method to optimize costs of building test collections when there is large number of topics and documents. Their suggestion is to use a resource allocation method to build a low-cost and reliable test collection. In their method, they considered prioritizing topic-document pairs when there is limited budget for IR evaluation. They specifically include a cost constraint within the optimization and formulate the optimization problem using existing algorithms for finding an overall optimum result (Hosseini, Cox, & Milic-Frayling, 2011).

In another study, Sanderson and Zobel (2005) examined the reliability of measuring IR system’s performance when using a large topics set. While exploring the consistency of an evaluation measure to assess the effects of using statistical significance tests, which are commonly used to verify the results of evaluation experiments, they compared MAP and P@10 considering the assessors effort required to calculate those measures. They stated that constructing test collections with more number of topics while performing fewer number of relevance judgment for each topic can be more effective than assessing the relevancy of lots of documents while evaluating systems over a few number of topics (Sanderson & Zobel, 2005).
3.7. Alternative methods to pooling

In this section, the following low-cost IR evaluation techniques: (1) without relevance judgments or (2) relevance judgments generated using alternative methods such as search results retrieved by a single effective retrieval system, will be discussed.

Relevance assessment is a critical point in retrieval evaluation because of the nature of how judgments are created. As mentioned, it incurs great costs to the evaluation process. In addition to the efforts to reduce the number of required judgments, researchers are seeking low-cost techniques that may substitute this task. One of the early attempts to replace human relevance judgment was the pseudo-relevance judgments approach which was proposed by Soboroff, Nicholas, and Cahan (2001). This approach was based on randomly assigning relevant documents to topics without performing any judgment. Their observations suggest that this random selecting of documents as being relevant, can obtain a fairly good approximation of relevance judgment set and reliable system rankings can be produced based on it. They used some ad-hoc TREC data to run the experiments and compared the results. The work of Soboroff et al. (2001) was a motivation for many IR researchers. As an example, Carterette and Allan (2007) ran an experiment to design and test an algorithm which ranks retrieval systems without relevance judgments. In the proposed method, $\varepsilon$MAP (expectation of MAP) is calculated for each topic $t$ in topic set $T$ as follows (Carterette & Allan, 2007):

$$\varepsilon\text{MAP} = \frac{1}{|T|} \sum_{t \in T} E[AP_t]$$

where expectation of $AP$ or $E[AP]$ in (5) is calculated from Eq. (6):

$$E[AP_t] = \frac{1}{p_t} \sum_{i=1}^n \left( \frac{1}{r(i)} p_i + \sum_{j=1}^{i-1} \frac{1}{\max(r(i), r(j))} p_ip_j \right) + \varepsilon$$

For document $i$, $p_i$ is relevance probability, $r(i)$ is rank, $n$ shows the total number of documents in the corpus and $\varepsilon$ is the error. Each topic gives a MAP that can be used to rank the systems with no judgments. To calculate $E[AP]$ the probability that a document $i$ is relevant has to be computed. If simply used, a constant value for $p_i$, when there is no judgment available, $\varepsilon$MAP for all systems would be the same. So, they used an aggregation algorithm for estimating the relevance probability.

Even though IR evaluation initiatives are using partial judgments, verification of relevant documents is still a resource consuming process. In order to tackle the problem of generating systems ranking in the absence of relevance judgments, Spoerri (2007) employed the assumption proposed by Lee (1997), which expresses that retrieval algorithms tend to retrieve equivalent sets of relevant documents and non-equivalent sets of irrelevant documents. In the experiment, the random groupings of five systems from TREC data were selected and the overlapping structure to produce a systems ranking was calculated. Based on the findings, Spoerri (2007) suggested that the overlap pattern between multiple search results can be used to recognize the quality of systems by excluding the need to determine relevant documents.

In another experiment, Spoerri (2005) investigated the overlap pattern between search results of retrieval systems participating in the ad-hoc track in TREC data. It showed that the potential relevance of a document rises exponentially as more numbers of systems are retrieving it; this is called the Authority Effect. In addition, as more systems retrieve the same relevant document, the document is increasingly placed near the top of the systems’ lists (Spoerri, 2005). Even though this method does not seem to be complete, it can be used to rank systems while there is no relevance judgment sets available.

Accordingly, Shi, Li, and Wang (2010) adopted the method proposed by Spoerri (2005) by considering the problem that it was usually ruined because of similar retrievals. Shi et al. (2010) tried to solve this problem using the method of clustering. In their proposed solution, they used one system as a representative of all similar systems, which can considerably lessen the unwanted influence of similar retrieval results. They found that this method outperformed all the similar judgment free efforts to produce a systems ranking (Shi et al., 2010). On the other hand, Sanderson and Joho (2004) examined creating qrels based on system pools and query pools. For query pools they used several search results retrieved by only one retrieval system to make qrels instead of pooling the result of multiple retrieval systems as in the system pools. Based on their experiments, they claimed relevance judgments can be created based on a set of qrels from the runs of an effective retrieval system. So, both the pooling method can be avoided by using a ranked output of that specific system which then can be used to create the relevance judgment set for the collection (Sanderson & Joho, 2004). In this approach, they need not use the pooling method for document selection and only the runs of a single system will be assessed for relevancy. This approach can be used to create a test collection rapidly and with few resources (Sanderson & Joho, 2004).

Another approach to rank retrieval systems without pooling-based relevance judgments is through topic presentation of queries (Efron & Winget, 2009). In this approach each query is presented in a variety of formats since the same information can be represented in different terms. In their experiment, for each topic $t$, a small set of other aspects of $t$ is created and a single IR system uses this set to find the relevant documents for $t$. So, running each aspect of $t$ as a query against the document corpus, a wide range of relevant documents for $t$ will be retrieved by the system. Accordingly, much less human effort is required to create different aspects of a topic rather than performing relevance judgments. However, estimated system rankings based on this approach could be as accurate as ranking based on the traditional method of pooling (Efron & Winget, 2009).
As a final discussion topic, one of the most recent methods to replace the traditional judgment approach and reduce experimental cost is crowd-sourcing. Crowd-sourcing is the act of sending out multiple tasks to a crowd of people resulting in a productive solution instead of assigning task to a single worker (Howe, 2008). This method will usually decrease the costs and time of performing tasks which makes it a desirable substitute for relevance judgment by human assessors. Alonso and Muzzaro (2009) proposed the use of Mechanical Turk crowd-sourcing platform (www.mturk.com) supported by Amazon® to assess the relevancy of documents. They used some of the TREC topics about the space program in the domain of science and technology to conduct the experiment and examine the results with existing relevance judgment sets. Comparing the results achieved from the crowd-sourcing method and traditional relevance judgments revealed that, crowd-sourcing was sometimes even more accurate than expert assessors in recognizing relevant documents. On the whole, they introduced this method as an appropriate alternative for the current method of relevance judgment (Howe, 2008). As IR evaluation initiatives were motivated to perform experiments to examine the effects of crowd-sourced data in evaluation results, TREC started the Crowd-sourcing track in 2010. In addition, other evaluation initiatives were also interested in this approach. In an experiment conducted by Kazai, Kamps, Koolen, and Milic-Frayling (2011) the effectiveness of the crowd-sourcing paradigm for book search evaluation was examined. Investigation on the effects of crowd-sourced data in INEX book track showed it can be an appropriate choice for IR evaluation.

4. Conclusion

Several state-of-the-art techniques proposed by IR researchers for measuring IR systems performance were considered in this paper. As the review shows, many attempts were focused on designing experimental methods and formulating a metric for effectiveness to be used when there is a limited number of relevance judgments and there is a need to generate reliable and accurate evaluation results. Even though several techniques in relation to low-cost evaluation have been proposed and discussed, for the essence of evaluation experiments, a standard approach has not been defined as yet and there is a need for further research in this area. This is because, generally, in the area of information and computing sciences, it is difficult to establish an IR evaluation methodology that will be accepted widely by the IR community. Even if such methodology emerges, it would not hold for long and they become dated fast.

On the other hand, within the area of IR evaluation, there are some important issues that researchers could address, such as how to reuse the data or results of system evaluation from varying collections to obtain even more accurate results in future experiments. A lot of effort has been put in order to generate the test collections (including judgments) and to measure the systems. Hence, these precious resources such as human effort and time should be re-used through a clearly defined methodology.

As we all know, the test collections based IR experimental methodology has been successfully practiced widely within the IR community for many years. But, being in a dynamic environment of information and computing sciences, it may be nearly impossible for an IR evaluation methodology to be sustainable. There will come a time where the existing test collection methodology and approach taken in measuring systems may be obsolete. Perhaps, it is time to find for new ways in undertaking a large-scale IR system evaluation or even measuring different aspects of the system (not just depending on document relevancy). However, the transition may not be immediate and it will be a gradual process.

Acknowledgment

This research was funded through the University of Malaya High-Impact-Research (UM HIR) Grant Scheme.

References


